Comments on ET Docket 02-135

[Originally filed July 8, 2002, revised slightly to correct typos and poorly worded parts pointed out to me and add some references that were not available to me at the time I wrote it]

Comments for FCC Spectrum Policy Task Force on Spectrum Policy

David P. Reed, Ph.D.
1242 South Street
Needham, Massachusetts 02492

Email: dpreed@reed.com

In my comments below, I focus on key technological and architectural issues that should have a major impact on Spectrum Policy in the US and worldwide. This technical and scientific approach contrasts strongly with the non-technical, purely economic and regulatory approaches that have governed thinking about radio communications for the last 70 years or so.

The arguments and proposals are based on my training, knowledge and experience as a systems designer for the past 3 decades, including contributions to the original architecture of the Internet protocols, my experience as vice president and chief scientist in the personal computer industry (Software Arts and Lotus Development Corporation) for a decade, and my research and experience in the past decade on mobile and personal wireless data communications systems, while affiliated with Interval Research Corporation, the MIT Media Lab, as an advisor and investor in startup technology companies, and on my own initiative and with my own personal funds.

I have spent the last 10 years looking at the fundamental technological, architectural, and economic issues related to the evolution of networked systems. Many of the technical ideas cited here are based on the work of others, some are my own, and any flaws in the synthesis are, of course, my own.

Due to prior travel plans and the unusually tight deadline, I have had to write these comments while traveling with my family in Europe. I had little access to libraries and reference materials, so I hope the lack of detailed references to the literature is not too much of a problem. I will be back in my office soon, and happy to respond to requests for further information by email at the address above.

Overview

I argue in this note that the foundation of a sound economic and regulatory approach to managing radio communications in the US and worldwide cannot and should not ignore fundamental advances in the understanding of communications technology that have been developed in the last few decades. Those advances are just beginning to reach the point

where they can be fruitfully applied in the marketplace, at a time when the need for a huge increase in communications traffic is beginning to surge.

It will be crucial for the continued growth and leadership of the US economy, and for its security as well, to embrace these new technologies, and follow them where they lead, in spite of the potential negative impact that these technologies may have on traditional telecommunications business models. There is a "new frontier" being opened up by the interaction of digital communications technology, internetworking architectures, and distributed, inexpensive general purpose computing devices. This new frontier cannot be addressed by a model that awards the telecommunications operators exclusive rights (such as "spectrum property rights") that can be used to "capture" the value yet to be produced by innovators in underlying technologies[1] or applications.

No economic "ether"

My argument is based on a simple but crucially important technical fact: the useful economic value in a communications system architecture does not inhere in some abstract "ether" that can be allocated by dividing it into disjoint frequency bands and coverage areas.[2] Instead it is created largely by the system design choices – the choice of data switching architecture, information coding scheme, modulation scheme, antenna placement, etc.

The most important observation about the impact of systems architecture on economic value is this: there exist *networked* architectures whose *utility increases* with the density of independent terminals (terminals are end-points, such as cellular telephones, TV sets, wireless mobile PDAs, consumer electronic devices in the home, etc.) Network architectures provide tremendous gain in communications efficiency on a systems basis – I call this *cooperation gain*, because it arises out of cooperative strategies among the various terminals and other elements in a networked system. (It should be emphasized that cooperation gain is not available to non-networked systems at all). Cooperation gain is discussed below.

The argument for this is purely technical – so it must be considered in any economic or regulatory approach that attempts to provide for allocation of economic value in communications systems. I would claim that a proper "market based" approach (and for that matter any proper non-market based approach) needs to focus on creating

[1] New technologies such as spread spectrum, smart antennas, ultrawideband radio, and software-defined radios create more capacity that cannot be known accurately until there has been broad practical experience and an industrial learning curve that reduces their costs. The FCC has consistently tried to base regulation on accurate forward looking prediction of the economic value of new technologies and new services, but those predictions have been consistently wrong. That isn't surprising given that the value is established decades later.

[2] The confusion that led pre-20th century physicists to postulate a "luminiferous ether" which carried radio and light waves has persisted in the economic approaches that attempt to manage communications capacity as if it were an "ether". Just as Einstein pointed out, counterintuitively to most, that there need be no "ether" in formulating Relativity Theory, recent results in multiuser information theory show that counter to the intuition of spectrum economists, there is no "information capacity" in spectrum independent of the system using it.

meaningful competition to create value for the users of a communications system. What drives our economy is the drive to innovate by doing a particular function more efficiently, or by enabling a new function that was impossible to be done at reasonable cost.

My second crucial observation about the impact of systems architecture on economic value is this: our understanding of communications systems, their applications in society, and the underlying digital technologies are undergoing enormous technological change, and that change is far from reaching any fundamental limits. Moore's Law has provided us with circuits that have doubled in capability in almost every dimension every 18 months for the last 40 years. We are nowhere near the limits – this evolution can be seen to continue, as long as economic demand for its benefits is allowed to continue, for the next couple of decades at least. The Internet has similarly demonstrated that architectural innovation in digital communications is creating capabilities at an exponential rate that does not seem to be anywhere near a technical limit. As the costs of digital communications drop, new uses stimulating new demands continue to arise.

The largest category of new uses will be new categories of networked devices – and this category will experience growth rates much larger than traditional communications terminals. Those devices include personal devices (such as cellphones and PDAs) and shared appliance-type devices (such as display consoles, control panels, consumer electronics, office storage devices, etc). They will be densely deployed, owned by users, small businesses, and corporations, and will be deployed in an unplanned, ad hoc manner. An important characteristic of such devices will be their mobility. Devices carried by a user or carried in a vehicle will be mobile as a result of people moving, whereas appliance devices will be moved just as furniture is moved – as workplaces and homes are rearranged. Communications reconfiguration resulting from mobility will be the norm – it would be unacceptable to the owners of these devices to buy new devices merely because one crosses an artificial spectrum boundary.

A key aspect of the new demand is that the systems can and will be largely "user financed" unless regulation bars users from deploying new technology. The bulk of capital expenditure in networked systems, especially systems that need no cables or optical fibers will be borne by users rather than by "network operators" as is the case in the cellular telephone and radio/TV broadcast industry. This is an important economic benefit, because it allows for direct investment by interested parties, rather than indirect investment, connecting demand directly to supply. (In the near future, this is trivially true, because there is little available indirect investment capital for telecommunications systems, due to the well-known collapses of network operators).

The remarkable success of 802.11b (WiFi) radio systems is clear evidence of the growth of this category of new devices and applications. Without any significant marketing expenditure, 802.11b WLAN terminals have grown at a remarkable rate, even during a telecom recession. Similarly, the interest in applications suggested by Bluetooth PAN technology has been enormous as well, though the technology as delivered has been disappointing in its limitations.

Regulating systems undergoing enormous structural and technological change is particularly challenging because the nature of regulation is to restrict flexibility that may later turn out to be crucial to the economic success of systems. One well-known approach to retaining flexibility in the face of change and uncertainty is based on modular design. In the case of communications systems, a very successful design principle has been the "end-to-end" argument employed in structuring the basic Internet protocol architecture. I was a co-author of the original paper defining the "end-to-end argument", and a participant in key Internet design decisions (splitting TCP into IP, TCP and UDP, and the separation of naming addressing and routing) based on it. Because of the high rate of change and high uncertainty about future radio technology and applications, we need to apply the end-to-end[3] principle to spectrum allocation as well. In the discussion below I will explain this in more detail.

In the early part of the 20th century, the idea of allocating radio by bands allocated to fixed services made sense, for several reasons. First, the only known technical means for multiplexing signals on a shared "ether" was using tuned resonant circuits, which could handle only a small band of frequencies designed into the terminals. Second, there was a vast frontier of new, higher-frequency radio bands yet to be exploited. Third, the important applications for radio communications systems were for long-distance transmission, which, absent networking and repeaters, made a dramatic difference in the utility of various bands due to propagation differences. In implementing this system we may well have made important mistakes - economic thinkers, building on the early suggestions of Ronald Coase in his seminal papers on the FCC and the IRAC, have suggested that a better way to assign frequencies to the best use would be to create a market in frequencies that can be bought and sold. This indeed might have been an appropriate way to resolve competing needs for communications, in the early 20th century.

In the early part of the 21st century we face a very different, evolving, and unpredictable set of demands for radio communications, and we have learned a great deal about the implementation of radio systems and networked systems in the last decade or two. We have a lot more to learn. We should not tie our future needs to an economic approach based on the assumptions and limits of early 20th century radio technology, and we should be careful about predicting future applications of radio communications as mere extensions of currently mature applications of that technology (such as broadcast radio & TV, radio dispatching services, walkie-talkies, and radio-telephony).

Faced with the choice of looking backwards to correct historical errors, or facing a tidal wave of new demands and new technologies that we don't understand very well, it is tempting to focus on what we know. But I would argue that this is a serious mistake. It would be much better to create an approach that allows us the flexibility to enable the systems architectures and technologies that are just emerging, rather than to optimize a system as if it were in equilibrium.

---

[3] J.H. Saltzer, D.P. Reed, D.D. Clark, *End-to-end arguments in system design,* ACM Transactions on Computer Systems **2**, 4: 277-288, November 1984.

Two major new developments have made this revolution possible. These developments in broad terms have reshaped the way we must think about communication systems, and we must reorganize our regulations to embrace their benefits.

The first development is internetworking, and the second is digital signal processing.

Internetworking (on which the Internet is based) consists in understanding that information is independent of the medium that carries it, and can be represented in a universal digital representation – the bit. What the Internet has taught us is that we need not design communications systems for voice bits that differ fundamentally from systems for video bits – instead, by carrying all kinds of traffic over whatever links are available, we can achieve a high degree of efficiency, both technically and economically. Interoperation between networks removes unnecessary transaction costs, enabling new applications to reach economically viable scale without the overhead of purpose-built networks for each new application, and enabling existing applications to be improved in an upward compatible way while allowing legacy versions to coexist.

Digital signal processing is the use of extremely inexpensive and rapidly improving digital technology to handle all aspects of processing signals, including tuning, modulation, coding, and compression, among other functions. Since digital technology enables complex and adaptive algorithms we are able to approach closer and closer to the theoretical limits involved in manipulating and perceiving aspects of the physical world – in the case of radio, directly manipulating and sensing the electromagnetic fields that can be manipulated to carry information. The result has been a dramatic reduction in costs to implement efficient and adaptive techniques such as CDMA, spread spectrum, ultra-wideband radio, agile radio, power management, etc. At the limit, radio technology approaches the point where each radio is a "Cognitive Radio" that can sense its electromagnetic environment directly and modulate electromagnetic fields directly in time and amplitude.

A new approach to radio regulation

I believe that the combination of internetworking and digital signal processing must be fundamental to a necessary new approach to regulation of radio communications. Such an approach must also encompass rapid change in technology and applications, and rapid growth as well.

What would such an approach look like? How would we measure its effectiveness? These are the two crucial questions that must be addressed by regulators.

First, we must accept that radio systems should form networks. Networks are almost always more efficient than independent systems.

Second, we must accept that radio systems will be interconnected to non-radio communications networks. It is no longer sensible or possible to limit a particular application service to a particular medium of transport.

Third, we must recognize that in the not-too-distant future, all radio systems will be based on digital signal processing, and thus will approach "Cognitive Radio" capability. By cooperatively sensing and manipulating their electromagnetic environment, a network of software defined radio transceivers can adapt to their physical environment to match demand much closer to the capacity achievable by joint action of a group of radios.

This is new territory – not explored by existing theories of regulation.

The best example in recent history of such a system has been the Internet, yet it has two crucial constraints that the new radio technologies don't: need for cable deployed along rights of way, and a fixed switching infrastructure built around statically deployed cable terminations.

The Internet has already stressed the existing regulatory framework beginning to eliminate distinctions between telephony and content distribution, for example. Future radio regulation must deal with those issues, and in addition deal with the fact that radio networks can be assembled easily with end-user capital. That is, the crucial economic actors will be the hardware and software product companies that develop radio connectivity and software-defined radio protocols to the public. Like the PC industry, the control of modular interfaces, standards, and protocol evolution will be the key areas of competition to define services for users, rather than the current situation where competition focuses on operators because they bear the capital costs of system deployment.

Managing vendors of network components that will be formed into networks by users will be the role of any regulatory approach. Obviously it is important to make sure that those network components work together efficiently, and that joint and societal benefits be maximized. Where market forces will encourage efficient interoperation, there should be no need for government intervention, but where market forces can't work well, the government may need to step in to manage things.

But what needs managing? In the next few sections, I discuss some fundamental metrics of capacity that ought to stand out as key metrics. Before discussing how these can be managed, we must discuss some goals.

Capacity: Bits vs. Hertz

Confusing information capacity with bandwidth is a source of great confusion in discussing radio systems.

Shannon defined the notion of information by defining the unit of information as a bit. He linked the capacity of a radio communications channel in bits/second to its bandwidth

in Hertz by a theorem that showed that the limit of information capacity in simple point-to-point channel is proportional to the bandwidth of the channel in Hertz. In such very simple systems, bandwidth and information rate can be treated almost interchangeably.

More complex systems, however, have capacities that depend on other factors beyond bandwidth. To be clear, I recommend that it is crucial to avoid using the term bandwidth[4] when "information rate" is meant, and to measure communications capacity in bits.

Any regulatory system should be using measures based on bits per second, rather than Hertz.

Measures of effectiveness

There are many different measures of bit-delivery effectiveness for a networked communications system – ultimately the measure depends on actual applications' needs. However, from the point of view of terminals in a system, its usefulness is based in how many bits of user level information can be carried, and how far those bits are carried. A natural measure of systems capacity is what I will call *transport capacity*. The transport capacity of a system is defined as the maximum achievable *transport usage* among the terminals in a system. One measures the transport usage by adding up the transport usage for all messages delivered to their final destination during a particular time – where the transport usage is defined to be the number of correct bits in the message multiplied by the distance between the original source terminal and the ultimate destination terminal.

Other important measures of effectiveness also need to be considered in evaluating networked systems – e.g. end-to-end delay in delivering messages, and the flexibility to allocate capacity among competing uses. However, the transport capacity is an important figure of merit that captures systems effectiveness much better than does "spectral efficiency", because it takes into account distance.

Another important measure is *channel transport efficiency.* This measure is the ratio between transport capacity as defined above and the sum of the radiated energy added to the system by all transmitters. In a fully mobile system, this is economically important because it is the fundamental constraint on battery life. It is also important because radiated energy has other important impacts, e.g. on biological and electronic systems. Reducing the amount of energy to achieve a given transport capacity is an important factor.

[4] The common "techno-cultural" usage of the term "high bandwidth" (and also "broadband") systems to mean systems that have a high communications rate in bits/second is the primary source of this ongoing confusion. It is apparently "cool" sounding to use the terms incorrectly in this way, despite the fact that "bandwidth" and "information rate" are not at all the same. Perhaps if we create an honorific term for bits per second, as we have for cycles per second, this confusion can be culturally corrected. The obvious metric for bits per second ought to be "Shannon".

An ideal system architecture and economic approach would sensibly maximize both the transport capacity of a system and the channel transport efficiency, subject to the constraints of actual communications demand.

In practice, optimality cannot be achieved because of two kinds of constraints.

First, the actual demand cannot be anticipated or modeled – it depends on *extrinsic* factors. Experience with the Internet has shown that natural demand is bursty at all timescales, with little statistical smoothing effect. And growth in overall demand (likely exponential in nature, like that for semiconductor performance in Moore's Law) makes it quite difficult to use prior experience to extrapolate future needs beyond short time intervals.

Second, dynamically assigning capacity to fluctuating demand involves communication itself. This is the major *intrinsic* factor – achieving an optimal assignment of energy use in the system requires communication among the parts of the system, which itself uses more information. Clearly there is a tradeoff between optimality and responsiveness to changing demand that involves deciding how much communications capacity should be allocated to the overhead of capacity management.

Part of the communications overhead necessarily involves the cost of determining where to invest additional capacity to meet future demands for capacity, assuming capacity needs tend to grow predictably. It is this part of the underlying system architecture where "price signaling" would be useful to users and intermediaries involved in the system.

What we know and don't know about systems capacity

Until recently, digital radio communications networks have been rare and small, special purpose appendages to the wired networks. We have little experience with dense indoor data networks, nor with high performance, densely deployed outdoor data networks. Further, it has been assumed (without thorough evidence or analysis) that the as terminal density increases, interference will cause the overall transport capacity achievable to degrade.

Thus, the capabilities of networks to provide capacity that increases with the number and density of terminals has been a recent discovery. Remarkably, though theorists have been investigating this problem for a number of years, we don't have an answer to the following simple problem:

> Given N terminals distributed randomly throughout a fixed region (area on a surface, or volume of space), how does the maximum transport capacity that can be achieved among those terminals behave as a function of N.

Understanding this problem is essential to understanding the effectiveness of an architecture in creating economic value in the form of transport capacity.

Yet this is an important unsolved problem in multi-user information theory.

What we do know is that it is possible to achieve transport capacity that increases as N increases, with known network architectures.

For example, we know that with relatively simple, ad hoc repeater-based radio-only network architectures, transport capacity can grow as the square root of N, or $N^{1/2}$ when the individual stations are located on a plane or the surface of a sphere.[5] We also know that transport capacity of a repeater network can grow as $N^{2/3}$ when stations are deployed in a three-dimensional volume, like downtown Manhattan.[6] Transport efficiencies in such repeater architectures scale quite well – the total energy needed to sustain the increasing capacity remains constant, so the transport efficiency grows also as N increases.

Deploying a cable-connected (copper or fiber) access-point network (such as a cellular network[7]) with a constant ratio of terminals to access-points creates a network whose transport capacity can grow proportionally to N, and whose transport efficiency grows as $N^{3/2}$. The practical limit of transport capacity in such hybrid networks is due, not to spectrum capacity limits, but to the cost of installing and maintaining the access-point network. Using radio-linked access-points does not scale well, however, since repeater-based networks have much higher transport efficiencies when all radiated energy in the system is considered.

Neither of these network approaches is known to be optimal according to the two measures considered. In fact, at the present time, despite active research in the area of multiuser information theory over the past 20 years, there is no tight upper bound on the transport efficiency achievable by radio systems architectures as the density of such systems increases.

For example, systems based on space-time coding (e.g., BLAST) have been analyzed, and shown to have transport capacities that increase proportionally to the number of antennas.[8] This technique (which has nothing to do with the repeater-based architectures)

---

[5] See Timothy Shepard, *Decentralized Channel Management in Scalable Multihop Spread-Spectrum Packet Radio Networks,* MIT Laboratory for Computer Science Technical Report TR-670, July 1995; also Timothy Shepard, *A Channel Access Scheme for Large Dense Packet Radio Networks,* ACM Computer Communication Review 26, no. 4, October 1996; also P. Gupta and P.R. Kumar, *The Capacity of Wireless Networks*, IEEE Transactions on Information Theory **46**, 2: 388-404, March 2000.

[6] See P. Gupta and P.R. Kumar, *Internets in the Sky: The Capacity of Three-Dimensional Wireless Networks*, Communicatoins in Information and Systems **1**, 1: 39-49, 2001.

[7] Not all access-point network protocols achieve this goal. To achieve this scaling of transport capacity and transport efficiency, a cellular network must *actively* manage the power of both the access points and terminals, using minimum energy protocols. Protocols such as the 802.11 standards today do not use adaptive power management, nor do many current cellular network protocols fully minimize system power.

[8] G.J. Foschini and M.J.Gans, *On Limits of Wireless Communications in a Fading Environment when Using Multiple Antennas*, Wireless Personal Communications **6**: 311-335, 1998, Kluwer. See also A.L. Moustakas, H.U. Baranger, L. Balents, A.M. Sengupta, S.H. Simon, *Communication through a Diffusive Medium: Coherence and Capacity,* Science **287**: 287-290, 2000, and S.H. Simon, A.L. Moustakas, M. Stoytchev, and H. Safar, *Communications in a Disordered World*, Physics Today **54**, 9 (September 2001).

has been shown, when combined with repeating, to create systems that scale so that transport capacity scales with N in interesting cases.[9] Another result suggests that architectures that take advantage of the motion of terminals may have transport capacities that scale proportionally to the number of antennas.[10]

Cooperation gain and increasing returns from networks

What is clear from analyzing networked architectures is that as the demand for capacity increases, and as the density of terminals increases, adaptive network architectures that involve cooperation among all of the communicating entities create radio systems whose capacity can scale as demand increases.

Compared to systems of dedicated, isolated links, networks provide much more transport capacity at much greater transport efficiencies.

This phenomenon, which I have begun to call *cooperation gain*, creates major economic benefits.

It is also well understood that networked architectures can provide dramatic benefits in terms of flexibility of connection and adaptation to demand. The system-wide option value of flexibility in a network scales proportionally to the square of the number of nodes, according to the law popularly known as Metcalfe's Law. Similarly, the option value that accrues due to the ability to dynamically assign capacity depending on shifting demand can increase superlinearly as the number of cooperating nodes in a network. I call these network externalities *network optionality*.

The availability of cooperation gain and network optionality suggests that as the number of radio terminal nodes increases, and demand increases, the most effective architecture for radio communications deployment will tend to be a small number, perhaps one, of interoperable, loosely coordinated networks that evolves according to demand.

Interference, Noise, and Signal

The analyses referred to above suggest that the engineering notion of "interference" used in analyzing point-to-point and broadcast systems is not the best way to approach the analysis of systems that involve many terminals.

Instead of minimizing interference on a receiver-by-receiver basis, the architectures above maximize the useful information rate delivered by the system as a whole among all the terminals involved. The combination of what is traditionally called "interference" and that proportion of capacity devoted to coordination of transmission and coding

---

[9] P. Gupta, P.R. Kumar, *Towards an Information Theory of Large Networks: An Achievable Rate Region*, IEEE International Symposium on Information Theory, Washington, DC June 2001; submitted to IEEE Transactions on Information Theory.

[10] M. Grossglauser, David Tse, *Mobility increases the capacity of adhoc wireless networks,* Proceedings of IEEE Infocom Conference, April, 2001.

overhead causes any reduction of capacity. Tradeoffs between "interference" impact at different points in the system and between "interference" and overhead communications for coordination are possible – managing the overall tradeoff to maximize application value is a global process that cannot be decomposed into individual link-based requirements.

Focusing (as the current FCC regulatory approach does) on interference rather than on useful capacity and optionality tends to optimize the wrong attributes.

For one trivial, but illustrative example, when a link is not actually needed to transmit useful information, external signals impinging on that link's receiver do not in any way impede that link's capacity. However, the S/I ratio as seen for the link is either irrelevant, or tiny, depending on one's point of view. Counting the impinging signal as "interference" is clearly wrong, but the error can only be corrected by considering this on a whole systems basis.

Though this is an extreme example, less extreme versions of this error pervade the current regulatory policy of the FCC. For example, one measures the interference among television broadcast stations over all geographic locations in the "footprint" of a licensed station, rather than merely at the set of actual receivers that are "tuned in" to that channel at a particular point in time. Thus, even though there may be no receivers listening to a particular station at a particular point in space, the regulations impute a "loss of capacity" at that point based on a S/I ratio that can be measured at that point in space.

Similarly, in order to provide coverage of the maximum possible area, interference among broadcast signals is measured at the extreme boundaries of geographical areas. Since signal strength declines as the square of distance, or worse, this means that the bulk of the licensed region receives signals that are far more powerful than is necessary for the data rates delivered. This excess power necessarily behaves as "interference" to stations nearby, even stations in different bands.

Networked systems that transmit at low power, with dynamic directivity and repeating of signals are far more effective in terms of the amount of information delivered per unit of energy. The reason this is the case is clear – rather than directing energy mostly to places where it is not needed, the system optimizes itself dynamically to the information channels actually in use – if a receiver is not "tuned in" to "channel 7", it need not be receiving energy from that source that would be interfering with its attempt to "tune in channel 5".

As long as the regulatory process (including litigation and lobbying, and even secondary markets) focuses on defining interference without reference to the actual dynamic uses of systems, and as long as there is no incentive among radio transmitters to create cooperation gain as networks, there will be no economic means to gain these reductions in "actual" interference (as opposed to the current measures of "imaginary" interference). Instead, the arguments and tradeoffs among radio licensees will focus on whatever

measures the FCC uses, which are currently far from those that matter to communications users.

Power limits

The current FCC regulations are based on limiting radiated power from a particular antenna. Presumably this resulted because such a limit is easy to measure (especially when there are relatively few transmitters).

However as the density and capacity of radio networks increases, this is clearly the wrong structure for control. In low-density applications, high power systems are clearly useful, but as density increases, there is no need to use high radiated power to overcome environmental noise – instead one merely creates an "arms race" among non-cooperative systems to "outshout" each other.

Instead, fostering cooperation gain and rewarding systems that use the minimum necessary power to achieve the desired end-to-end bitrate is the approach that benefits all.

The overall system capacity of a collection of radios operating at a particular total power is not increased by doubling the power of each radio, or in most cases, by reducing the power of each radio by 25%. However, any one radio doubling its power creates a temporary gain in capacity relative to all others. Thus in the absence of any incentives for cooperation, natural competition for capacity on a link-by-link basis will not result in transport efficiencies.

Instead, mechanisms that enforce cooperation and minimum power will almost certainly need to evolve in order to create efficient and scalable capacity.

The FCC has created barriers to network interconnection

One of the most serious problems (from a technical point of view) with the current regulatory structure imposed by the FCC is that it actively blocks internetworking.

We have shown that by treating "bits as bits", cooperative network structures can achieve very high transport efficiencies and scalable transport capacities. And as noted above, networks create optionality that further increases their economic value.

Yet throughout the FCC regulation of radio there are three key, explicit kinds of restrictions that prevent the creation of networks, and a fourth restriction that limits effective networking is implicit in rules that restrict the use of security transforms.

The first kind of restriction found is that most bands and services bar "repeaters". Repeating signals is essential to cooperation gain of the form we note above. The historical basis for barring repeaters probably lies in the notion that one cannot easily define personal or corporate accountability for whatever "interference" is caused by a repeater station, and since regulation focuses on power limits, repeaters are likely to

operate at maximum power in the most optimal location to overwhelm all other transmitters.

This bar on "repeaters" bars nearly all kinds of networking within bands. Recent actions by the FCC show that these kinds of restrictions are typically added without much thought to newly licensed services. For example, the recent R&O regarding UWB radio services bars use of repeaters, and the recent licensing of a 1.2 GHz service similarly bars repeaters.

The second kind of restriction that inhibits internetworking is that interconnection between networks operating in different bands is barred in many cases, or restricted by specification of particular architectures. For example, "phone patches" have traditionally been barred, and frequency translators and remodulators are highly restricted.

The third kind of restriction that inhibits internetworking is that specific bands are reserved for different "services" – that is, different kinds of applications and different kinds of content. Historically, the inability inexpensively to create receivers that can be easily reconfigured to use different modulation techniques and different frequency bands may have justified this requirement to maximize economic benefits to users. However, low cost frequency agile digital receiver designs now exist and are getting much cheaper. There is very little economic benefit from static channel and modulation choices in terms of receiver cost, while much larger benefits would inhere from internetworking that allows dynamic assignment of capacity to connect to services as needed.

As Nobelist Ronald Coase pointed out in his 1959 paper entitled "The Federal Communications Commission", one impact of statically associating services to bands may well have always been to restrict the capacity available for certain kinds of communications, thus providing "cover" for US Federal regulation of the information content transmitted using radio.[11] In this paper I will not focus on this legal concern, which is outside the scope of my technical arguments, but it seems clear to me that there is a strong First Amendment argument against any regulation that unnecessarily limits constitutionally protected speech over radio. Since much greater information capacity would result from internetworking and dynamically adaptive radio architectures, it would seem that barring internetworking and adaptive digital radio is not only economically inefficient, but also legally unconstitutional.

The fourth, implicit, barrier to internetworking is the requirement that the content of communications be exposed in an insecure manner to all who can receive the signal. Internetworking does not require that content be obscured while in transit, but effective use of intermediaries to carry messages on one's behalf requires that one can trust those intermediaries not to alter or expose those messages in a manner not authorized by the sender or the receiver. In every form of communications transport other than radio,

[11] Ronald Coase, *The Federal Communications Commission*, International Journal of Law and Economics, 1959. See also, Ronald Coase, *The Interagency Radio Advisory Committee*, International Journal of Law and Economics.

messages are wrapped at the source, and only a limited amount of information that is needed by the intermediaries need be exposed, such as address, value, priority, etc.

In particular, the key "rights" in an internetworking system is the right to obtain carriage for a message over some set of network elements between the source and destination. These rights are easily implemented by tags placed in messages that provide unambiguous and unforgeable indications of the authority requiring the carriage of information. A simple example of this concept is a label on a message that indicates membership in a group that has made prior arrangements to carry each other's traffic for mutual benefit.

Modern digital communications security techniques are known which can reliably detect modifications to messages and prevent exposure of content by intermediaries that are not fully trusted. I refer to these techniques collectively as *security transformations*, and they include such ideas as digital signatures and end-to-end encryption of messages. Similarly, dispersing the energy of a signal across a wide band (using spread spectrum, space-time coding, and UWB techniques) can provide security transformations by making it difficult for an individual intermediary to understand or modify the fully dispersed message.

By providing end-to-end insurance that messages are not read or undetectably modified in transit, use of security transformations encourages competition among various intermediaries in the network on terms that benefit the end users – intermediaries can only read those portions of messages that are needed to deliver the message, and failure to deliver can be traced back. This reduces the need for specially "vetted" intermediaries and reduces the need for regulation of intermediaries – any particular intermediary may not be able to act alone in exposing or modifying a message.

In summary, then, there are four ways in which the structure of current radio regulations prevent achieving the benefits of internetworking. Each of these must be gradually removed from the structure of regulations, as follows.

First, barriers to repeating of signals must be eliminated from regulations. This must be done in conjunction with moving away from power limits as the means for controlling interference. For example, one might allow repeaters as long as the total energy emitted in the transmission of a message from source to destination is less than would be the case without a repeater involved. Implicitly this calls for active power management.

Second, barriers to interoperation between bands need to be phased out. Messages sent in one technique on one band should be allowed to be copied and retransmitted on other bands.

Third, limits to the type of content that can be transmitted in a band, modulation scheme, or systems architecture need to be eliminated.

And fourth, barriers to the use of security transformations need to be modified or phased out as internetworking is phased in. While there may be public policy needs that call for the ability of law enforcement authorities to intercept and trace communications, these needs may be satisfied by means that do not enable untrusted non-governmental intermediaries the same level of access.

Clearly these steps amount to a dramatic shift in the current structure of technical regulations, and full implementation of these steps will require that the historical legacy of systems and economic structures based on older technologies be gradually replaced by more efficient functional approaches.

The new radio communications regime

Uncertainty about what technologies and architectures will be optimal in the long run cannot be the basis for delaying innovations that will enable rapid growth of needed capacity and economically beneficial optionality.

Instead we must create a process that will enable competitive technology and architectural developments that move us closer and closer to a new, highly scalable and user-financed digital radio communications capability.

The two key elements involved will be phase-out of obsolete and inefficient legacy architectures, and enabling of commercial incentives for new architectures, at first for new applications and eventually for legacy applications.

The key "right" that users of radio networks will pay for is the right to pass messages over some intermediate networks in order to deliver the messages to the desired destination. Such a right can be represented by labels using digital signature techniques, and such rights can be traded in markets that create price signals for additional network investment. The details of these techniques are not difficult to develop – however the optimal structure of markets, like those of any other institutional market systems need to evolve as the applications and economies around those markets evolve.

As I have pointed out earlier in this paper, internetworking of adaptive digital radios will be the future of all radio communications. However, the strongest need for these systems will be in new applications areas, such as the short-range, dense networking of new devices to be carried on the person or distributed throughout personal and business spaces, and the creation of competitive "local access infrastructure" where physical rights-of-way or cabling have been extremely expensive[12] compared to radio-based solutions.

[12] Areas that are either remote (for example sparse rural populations) or where rights to deploy cables are accompanied by very high costs (impact on streets or requirements of access to building cableways) are good examples of cases where user-financed digital radio networks can provide very effective solutions for little or no capital cost. For example, 802.11b networking elements can be used today in place of broadband cabling; with adaptive power management and scalable protocols, similar elements can be deployed in order to provide solutions that can be financed where the incremental cost of adding to the

The combination of the decline in cost of individual radio elements, coupled with the desirability of building these systems on a pay-as-you-need-to basis suggests that most, if not all, of the capital cost involved in building these networks will be user-financed. That is, the users will buy the equipment and software that they need to build the networks as they need capacity. Where these radio networks interconnect with networks that have a high fixed-cost (such as wired networks), arrangements for carriage can be financed by charging for "rights" to transfer messages between wireless and wired networks.

Like any economic systems, however, these network structures create opportunities to "game" the system. Bad actors will still be able to "jam" communications, though attempting to cause widespread and sustained disruption has a much higher cost than it does in today's radio systems, due to the decentralized and adaptive nature of these networks. Just as today's Internet and the decentralized market economy in the US can be disrupted locally, but are globally resilient, so will these networks. Detection of disruptive actions and punishment of bad actors will still need to be a collectively shared cost.

A possible roadmap to the new regime

Having described the desirable end-state, here is a rough outline of the steps I believe are needed to get there.

First, we need to define what I call "the neck of the hourglass".[13] This is a communications protocol that is independent of the potential underlying transmission architectures that enables internetworking of radio systems. Like the Internet protocol layer called IP, it should be as simple as possible, while allowing the expression of the desired communications functionality. IP was a first draft, and is not sufficient or appropriate for radio internetworking. But at this point in time we know enough to make a good first approximation to a "radio IP" that can be used to begin the process, which can then be evolved in an upward compatible manner, as IP itself has been. "Radio IP" involves two aspects of a standard – first a standard for describing a broad and compatible set of modulation/demodulation techniques within a band, and a technique for sharing code in a high-level language that can be used to deploy those techniques on a range of software-defined radios.

Second, we must provide for major evolutionary steps. I suggest that the best way to do this is to open up capacity for this new regime in a sequence of phases. In each phase, an economically meaningful amount of bandwidth should be opened up for use by the next generation of software-definable radio hardware, software, and protocol technology. The "neck of the hourglass" protocol will be used, but periodic, and predictable releases

---

system is a small amount per user, rather than a high fixed capital cost that can only be justified where there is a sufficient density of subscribers..

[13] The "neck of the hourglass" refers to the famous concept of the "hourglass model" named by David Clark, where the IP protocol is the common point of interaction among a plethora of application-specific protocols built on top of it, and a plethora of technologies, architectures and protocols related to transporting bits underneath it. Creating the IP layer was a crucial application of the end-to-end argument.

of new tranches of spectrum will allow room for new innovations that will be enabled by the march of technological innovation--i.e., theoretical discoveries, and physical techniques etc. that may not be compatible with the initial techniques deployed in the first tranches of spectrum.[14]

Third, we must provide strong incentives for efficient use of the shared medium by protocols that adaptively manage power and achieve significant cooperation gain. The availability of joint benefits to all users, in terms of transport capacity, transport efficiency, and various kinds of optionality, is the primary incentive, of course. A user investing in new equipment, or considering becoming a participant in a particular new network must see immediate benefits. Here the strongest incentive is likely to be the ability to interconnect with existing high-speed wired and wireless networks that have broad connectivity – in particular the Internet as a whole. We can use communications rate and ability to access to the full Internet as means to penalize inefficient stations. Intermediate nodes that detect a sender is transmitting with more energy than necessary can drop messages with a probability proportional to the degree of excess energy, for example.[15] Such means would not be available to discipline terminals that do not choose to participate in the benefits of internetworking protocols, but radiate in the band. However such nodes will be a minority and their impact can be mitigated by means that depend on their minority status.

Fourth, as these adaptive network systems begin to mature, legacy applications will naturally begin to migrate to more efficient and flexible networks. Just as text messaging has largely migrated from postal mail and fax technology to email, and direct marketing has begun to migrate from mailed catalog shopping to online e-commerce, so traditional capabilities such as mobile telephony and music distribution will migrate to these new adaptive digital radio networks, because they are more efficient and flexible. At some point in time, the legacy architectures will be occupying spectrum better used in new ways. At that point, it may be necessary to retire old spectrum to be reused in new ways. I suggest that the best way to accomplish this retirement is to recognize that new technologies can begin to overlay existing spectrum in an agile manner before the existing systems are fully retired. Since the new technologies are "software defined" they can be deployed first in elements of geography and spectrum that are not actually being used, even though licensed for use. Phase-out of existing licensees can thus be done gradually, without a "flag day" and depending on natural obsolescence of existing equipment. The government can provide incentives by modifying the licenses to block

[14] The primary benefit of the end-to-end argument in economic terms occurs when market uncertainty is very high: the modular boundary that is encouraged by the end-to-end argument allows for phased changes while preserving most of the investment. The purpose of rolling over spectrum to new uses is to create the opportunities to roll out innovations and new applications where only the major modular boundaries (such as the hourglass neck) stay the same. Mark Gaynor in his recent Ph.D. thesis has built on the work of economists Baldwin and Clark to show that the end-to-end argument is most valuable in times of high market uncertainty and technological change. See the book by Carliss Baldwin and Kim Clark, *Design Rules*, MIT Press, 1999, and Mark Gaynor, *The Real Options approach to standardization*, Proceedings of the Hawaii International Conference on Systems Sciences, 2001.

[15] This idea of mine is inspired by the congestion control mechanism developed recently for TCP-like traffic on the Internet, called RED (for Random Early Drops), where dropping packets proportional to the degree of overload is used to discipline flows without requiring detailed knowledge of the flow sources.

new deployment or new purchases of legacy devices, and upgrades of existing facilities beyond the footprint of prior uses.

This roadmap is intended to be suggestive only. Many details remain to be worked out, and must be worked out in conjunction with our unfolding understanding of the underlying techniques that become available, and of the unanticipated applications that become possible as a result of rapidly scalable radio networking capacity.

Why we must start now

I believe that the exponential growth in demand for new capacity generated by new applications will, like any exponential, continue to accelerate. Though we only see the outlines of this new demand for digital communications via RF, it will be upon us very quickly.

We cannot afford to design the optimal answer before we begin to experiment commercially with systems that achieve cooperation gain through software-based adaptable radio systems.

Internationally, many nations are poised to begin much more quickly than we are to experiment with such systems. Their radio infrastructures are much more flexible and less legacy-based than those of the US. Though they have tended to follow our US regulatory structure, recent experience shows that the rate of uptake of radio innovations (such as 802.11b, for example) outside the US can be much higher than our domestic uptake.

The open and largely user-financed framework of the Internet architecture has indeed changed the face of wired communications over a period of 25 years. I believe that we are at the very beginning of a similar 25-year revolution. But we cannot create that revolution without changing the fundamental structure of radio regulation to focus on adaptive, digital, internetworked radio systems, financed by users rather than operators.

Appendix: Why Secondary Spectrum Markets Are Not a Good Solution

I have written this paper in an attempt to propose a constructive approach to new spectrum policy. As such, a critique of the idea of secondary spectrum markets is really not my main point.

However, I am very concerned lest the proponents of secondary markets as a solution to the spectrum shortage succeed in their quest to “propertize” spectrum.

The basic reason is that the bulk of the value created by adaptive networks and cooperation gain comes from the voluntary actions of users (deploying new devices, protocols, and systems.) That value, already paid for by the users, devolves to those users directly in the form of useful applications of the communications system. There is little or no value that is retainable by “spectrum owners” when the capacity of the spectrum

increases merely by adding more users who pay for their own physical and software capital. The only way for a spectrum owner to create a return, when capacity grows as the user participation grows, would be to create an unnecessary scarcity of communications capacity; that is, artificially raising prices by blocking users from using their own investments in hardware and software capital freely.

A different way to explain the same point is to assume that the state government holds the entire future value (in an economic sense) of all possible radio applications in its hands in trust for the citizens.

If we assume that the capacity (or utility) of all of the available spectrum is maximized by cooperative, user-financed sharing of the spectrum by a user-financed, adaptive digital network, what price should the government ask for transferring the right of development from its users to a collection of private holders?

Since all of the benefit will still be generated by investments of the users in equipment and in applications of the network, the spectrum holders' future investment need be nearly negligible, once they own the spectrum. But the potential return depends entirely on the exponentially increasing demand owing to future users, which the owners as a cartel will have the unlimited right to block.

Any finite price paid for spectrum rights as a whole is clearly too low.

Or to put it another way, if my technical argument about how value is created in radio networks by user financed investment is correct, then by selling spectrum rights to private holders for all time at *any* finite price, the government is *not* encouraging capital investment in the economic development of a resource (the usual argument for privatization.) Instead, it is selling out the future value of a resource best developed by individual citizen/users to a group of arbitrageurs -- whose best payoff is achieved by making no investment at all, while waiting until the public has to buy it back at a guaranteed substantial premium.